Annotated Transcription Gujarat High Court's Online Proceedings

Kunal Panjwani
DAIICT
Gandhinagar
202018001@daiict.ac.in

Vanditha Vinod
DAIICT
Gandhinagar
202018003@daiict.ac.in

Nihar Shah
DAIICT
Gandhinagar
202018014@daiict.ac.in

Akash Gupta
DAIICT
Gandhinagar
202018015@daiict.ac.in

Aakanksha Shah
DAIICT
Gandhinagar
202018026@daiict.ac.in

Yagn Purohit
DAIICT
Gandhinagar
202018035@daiict.ac.in

Sharvari Gokhale
DAIICT
Gandhinagar
202018038@daiict.ac.in

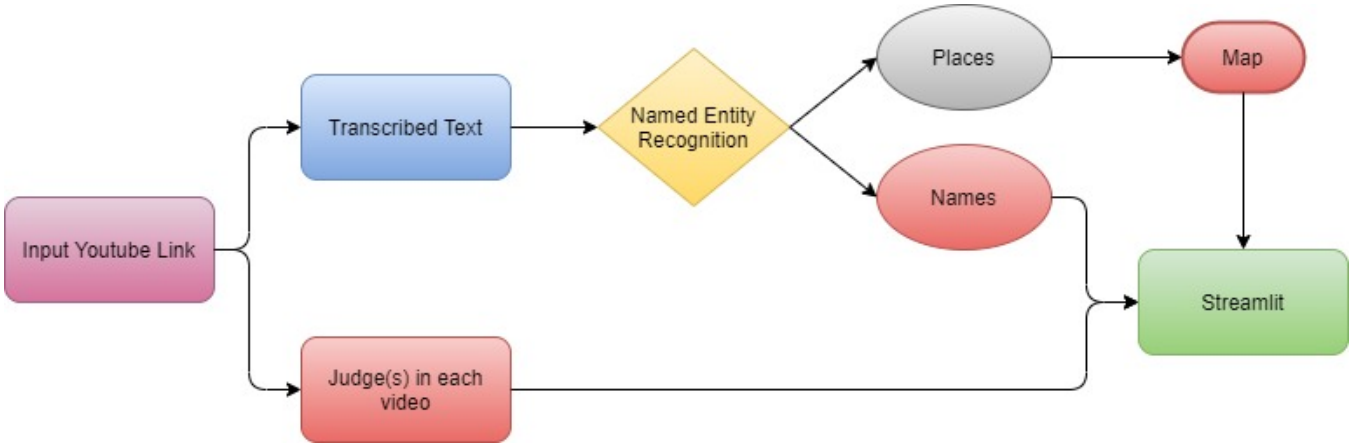

Figure 1. Project Pipeline

Abstract

In this project, we have transcribed the online proceedings of the Gujarat High Court, and have extracted names and geopolitical entities, using Named Entity Recognition and Part of Speech tagging, from the transcriptions. We also extracted the judge(s) name(s) from the description of each video. We then plotted these geopolitical entities on a map to give an idea about the spread of the proceeding.

Keywords: NER, POS, Gujarat High Court, transcription, annotation

1 Introduction

Due to the ongoing pandemic regarding COVID-19, people have opted to do their regular work at home (online), to maintain social distancing. Therefore, hearings of the court cases are also being conducted virtually. Like other high courts, the Gujarat High Court[6] has switched to virtual hearings after the COVID lockdown was put in place in March 2020. But, the Gujarat High Court on 26th October, 2020 became the first high court in the country to live-stream its court proceedings on YouTube in order to enhance access to courts, especially during the COVID-19 pandemic.

Owing to the fact that a large amount of legal data is being uploaded, the natural tendency is to summarize, and extract meaningful & significant information from these proceedings.

Our project implements some ways in which we can extract such information to some extent. It gives a vague idea regarding the judges & lawyers present that day, and the geographical spread of the cases that took place on that particular day.

2 Dataset

For the dataset, we used 14 YouTube videos of the proceedings uploaded by Gujarat High Court[3], on the basis of their dates. We converted these videos to audios, and then transcribed these proceedings using Otter[9], which is an online service for transcribing audio files. These transcriptions are the base of our dataset. We then used these transcribed texts for further annotation.

3 Methods

3.1 Extracting names of judge(s)

We noticed that the judge(s) names are present in the description of every video, following a certain naming pattern. So, we created a pipeline using youtube-dl[1], an open source download manager for YouTube, that takes a video's URL as the input and gives the text in the description. Using this pattern, we extracted the names of the judge(s) present on that day.

	Judge(s)
1	HONOURABLE MR. JUSTICE VIKRAM NATH
2	HONOURABLE MR. JUSTICE VIKRAM NATH
3	HONOURABLE MR. JUSTICE J. B. PARDIWALA

Figure 2. Name of the Judge(s) present on 24th December,2020.

3.2 Named Entity Recognition

We pre-processed the transcribed text. Using NER, we classified the text into various naming entities. From this, we extracted names of people & geopolitical entities. There is a high chance that the names in the 'Person' entity are of the lawyers.[7]

3.3 Geographical Spread

We plotted the geopolitical entities on a world map, by extracting the latitude & longitude of each entity using Nominatim API[2], which is a tool to search OpenStreetMap (OSM) data by name and address. An example of the same, which was done for 24th December, 2020[4] proceeding is shown in Figure 5. The density of the points represents how frequent the location has been mentioned during the proceeding.

	Names
11	Mehta
12	Ramanathan Singh
13	Salim Mehta
14	Huang Cha
15	Jim
16	Singh
17	Singh
18	Beta
19	Kanika Weaver
20	Charlene Mehta
21	Mehta

Figure 3. List of Names

	Places
16	Empire
17	Empire
18	DC
19	Jameson
20	Nepal
21	Oregon
22	Delhi
23	Bombay
24	Delhi
25	Delhi
26	Delhi

Figure 4. List of Places

Figure 5. Geographical spread of the proceeding on 24th December 2020.

3.4 Interface

We created an interface where we select the date we want, and we can view every detail mentioned above for the proceedings that took place on that particular date. We can view the judge(s) names, names of the people involved (lawyers, witnesses, etc.) and the geographical spread on the world map. We have the option to view which details we want.[5]

4 Use Cases

Keeping aside the accuracy, automating tasks such as transcriptions and annotating names, places and organizations can help in not only saving a lot of man power, but it can also speed up the process. The map can help in understanding and visualizing the spread of the proceedings around the world. The names of people and places involved in the proceeding makes it easy to summarize it and get an overview.

5 Limitations

At first, we wanted to create a CRF model[8] to create legal annotations on the proceedings, but the domain knowledge required regarding the legal sector was not available and hence manually annotating the data (which was supposed to be our training data for the model) for other legal entities apart from names was not possible. Also, the transcription platforms were not very accurate leading to some discrepancies in transcriptions.

6 Future Scope

Taking assistance from domain experts will help in generating legal annotations for the data suitable to a legal format. The interface created can also be extended and tweaked for displaying additional information. A better pipeline can be created to get more accurate transcriptions of the proceedings which will in-turn improve the accuracy of the named entities. Also, automating all these tasks is necessary if we are looking to apply it in real life.

7 Conclusion

Currently, due to the ongoing pandemic our dependence on technology has increased drastically. The more we can automate important jobs, the more we can save time & human resources. This project has hardly scratched the surface of how we can help the High Courts not only in Gujarat, but all the states of India.

The current project, i.e., extracting the named entities and the geographical spread, just by entering the date of the proceeding, is a good start for summarizing court proceedings.

8 Acknowledgments

We would like to thank Prof. Prasenjit Majumder, who guided us throughout the project. He provided us with many insights and has given us ample amount of time whenever it was required.

We would also like to thank our classmate Preet Patel, who has helped us tremendously in the transcription of the proceedings.

References

[1] 2006. *Youtube-dl is a command-line program to download videos from YouTube.com.* https://youtube-dl.org

[2] 2012. *Open-source geocoder with OpenStreetMap data.* https://nominatim.org

[3] 2020. *Gujarat High Court Youtube Channel.* https://www.youtube.com/channel/UCZoBFtdYPm8tBfGDzf4jsUg

[4] 2020. *Youtube Link of the proceeding of 24th Dec, 2020.* https://www.youtube.com/watch?v=6iVZPuR9I2Q

[5] Amanda Kelly Adrien Treuille, Thiago Teixeira. [n.d.]. *An open-source Python library that to create custom web apps for machine learning and data science.* https://streamlit.io

[6] Gujarat High Court. [n.d.]. *Gujarat High Court Website.* https://gujarathighcourt.nic.in/

[7] Susan Li. 2018. *NLP with Python.* https://github.com/susanli2016/NLP-with-Python/blob/master/NER_NLTK_Spacy.ipynb

[8] Kripabandhu Ghosh Saptarshi Ghosh Adam Wyner Paheli Bhattacharya, Shounak Paul. 2019. Identification of Rhetorical Roles of Sentences in Indian Legal Judgments. *arXiv* (Nov. 2019). https://arxiv.org/abs/1911.05405v1

[9] Simon Lau Sam Liang, Yun Fu. [n.d.]. *Online Transcription Service.* https://otter.ai